same as 10/033,823 filed 4 Oct 2002

PGPUB-DOCUMENT-NUMBER: 20040068514

PGPUB-FILING-TYPE: new

DOCUMENT-IDENTIFIER: US 20040068514 A1

TITLE: System and method for biotechnology information access and data analysis

PUBLICATION-DATE: April 8, 2004

INVENTOR-INFORMATION:

NAME RULE-47	CITY	STATE	COUNTRY
Chundi, Parvathi	Cupertino	CA	US
Collins, Patricia	Mountain View	CA	US
Graham, Simon	Palo Alto	CA	US
Vailaya, Aditya	Santa Clara	CA	US

US-CL-CURRENT: 707/102

ABSTRACT:

Systems and methods for database searching and data analysis with simultaneous, unified access to multiple heterogeneous data sources with effective reuse of user search session information for data analysis. The systems comprise a data source containing at least a partial copy of at least two public databases, at least one search program module operatively coupled to the data source and configured to carry out a search of the databases in the data source according to a user query, a data mining module operatively coupled to the data source and configured to provide for clustering of search results or documents from the user query and a user interface program module operatively coupled to the search program module and the data mining module, the user interface program module configured provide a visual interface for creating the user query and viewing the search results.

---------- KWIC ---------

Summary of Invention Paragraph - BSTX (12):

[0010] In certain embodiments, the data mining module is further configured to identify search results or documents according to a selected reference. The data mining module may also be configured to form clusters of related search results or documents according to an **unsupervised clustering** procedure, and may be is capable of preparing a single list of all search results or documents retrieved independently of the **unsupervised clustering** procedure. The data mining module may further be configured to assign a relevance score to the search results or documents based upon a frequency of terms from the query that appear within each of the search result.

Summary of Invention Paragraph - BSTX (13):

[0011] The **unsupervised clustering** procedure performed by the data mining module may employ a group-average-linkage technique to determine relative distances between the search results or documents. The group-average-linkage technique employs an algorithm for determining a proximity score that defines relative distances between the search results, the algorithm comprising

Summary of Invention Paragraph - BSTX (16):

[0013] The methods of the invention comprise, in general terms, providing a data store containing at least partial copies of at least two public databases, formulating a query by a user, submitting the query uniformly to each database in the data store, fetching search results or documents based on the query, and forming clusters of related search results or documents by a data mining module according to an **unsupervised clustering** procedure. The methods may further comprise displaying the clusters of related search results on a user interface and/or storing the clusters of related search results in a user data store. The methods may additionally comprise storing at least one user action, associated with submitting of the query, in the user data store. In certain embodiments, the methods may comprise defining a reusable query script and storing the query script in the user data store, and identifying repetitive user actions and storing the repetitive user actions in the user data store.

Summary of Invention Paragraph - BSTX (17):

[0014] In some embodiments of the invention, the methods may comprise identifying search results by the data mining module according to a selected reference. The methods may additionally include preparing, by the data mining module, a single list of all search results or documents independently of the **unsupervised clustering** procedure, and assigning a relevance score, by the data mining module, to the search results based upon a frequency of terms from the query that appear within each of the search result. In certain embodiments, the forming of the clusters of search results may comprise employing, by the data mining module, a group-average-linkage technique to determine relative distances between the search results. Employing the group-average-linkage technique may comprise employing the above algorithm for determining a proximity score that defines relative distances between the search results.

Detail Description Paragraph - DETX (19):

[0037] The data mining module 32 forms clusters of related search or query results according to an **unsupervised clustering** procedure and displays the clusters of related search results on the user interface.

Detail Description Paragraph - DETX (20):

[0038] The data mining module 32 is further capable of preparing a single list of all search results retrieved as **raw data,** independently of the **unsupervised clustering** procedure, after eliminating results not reachable via the web. The data mining module 32 assigns simple relevance scores to the search results based upon a frequency of terms from the query that appear within each document. The search results are then listed in the single list in an order ranging from a highest to lowest simple relevance scores.

Detail Description Paragraph - DETX (23):

[0041] Still further, the data mining module may process the **raw data,** independently of the **unsupervised clustering** procedure and the single list generating procedure, to categorize the search results so that each search result is assigned to one of a predefined number of categories. A list of words may be provided for each of the predefined categories wherein the words in each list are particular to the respective category. The data mining module 32 compares the words in a particular list to a document to be characterized to determine whether the document is classified in that particular category. Upon completion of categorization, the search results are also displayed in a categorized format to the user interface.

Detail Description Paragraph - DETX (26):

[0044] Well known **unsupervised clustering** techniques, such as the

group-average-linkage clustering algorithm ([A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, 1998, Prentice Hall, Englewood Cliffs, New Jersey]) can be used to determine relative similarities between documents. A particular example of a group-average-linkage technique that may be employed uses the following algorithm for determining a proximity score S.sub.ij that defines relative distances between search results:

Detail Description Paragraph - DETX (29):
[0046] The word "term" used above corresponds to a word in a search result (stop words may or may not have been removed from the search results). Stop words are list of words that occur very frequently in search results (such as common English words) and are deemed as insignificant in identifying similarities between search results. The use of this **unsupervised clustering** technique is also described in U.S. patent application Ser. No. 10/033/823 entitled "Domain Specific Knowledge-Based Metasearch System and Methods of Using" filed Dec. 19,2001, the disclosure of which is incorporated herein by reference.

Claims Text - CLTX (10):
9. The system of claim 1, wherein said data mining module is further configured to form clusters of related search results according to an **unsupervised clustering** procedure.

Claims Text - CLTX (11):
10. The system of claim 9, wherein said data mining module is capable of preparing a single list of all search results retrieved independently of said **unsupervised clustering** procedure.

Claims Text - CLTX (13):
12. The system of claim 9, wherein the **unsupervised clustering** procedure performed by said data mining module employs a group-average-linkage technique to determine relative distances between said search results.

Claims Text - CLTX (15):
14. A method for data access and data analysis, comprising (a) providing a data store containing at least partial copies of at least two public databases; (b) formulating a query by a user; (c) submitting said query uniformly to each said database in said data store; (d) fetching search results based on said query; and (e) forming clusters of related said search results by a data mining module according to an **unsupervised clustering** procedure.

Claims Text - CLTX (22):
21. The method of claim 14, further comprising preparing, by said data mining module, a single list of all search results independently of said **unsupervised clustering** procedure.

Claims Text - CLTX (35):
34. The system of claim 25, wherein said data mining means further comprises means for forming clusters of related said documents according to an **unsupervised clustering** procedure.

Claims Text - CLTX (36):
35. The system of claim 34, wherein said data mining means further comprises means for preparing a single list of all said documents retrieved independently of said **unsupervised clustering** procedure.

Claims Text - CLTX (38):

37. The system of claim 34, wherein said **unsupervised clustering** procedure employs a group-average-linkage technique to determine relative distances between said search results.